Chapter 3. Transcriptomes and Proteomes

Learning outcomes

When you have read Chapter 3, you should be able to:

1. Define the terms ‘transcriptome' and ‘proteome'
2. Draw a diagram illustrating the modern interpretation of the genome expression pathway, indicating the main points at which genome expression is regulated
3. Distinguish between coding and non-coding RNA and give examples of each type
4. Outline the process by which RNA is synthesized in the cell
5. List the major types of RNA processing events that occur in living cells
6. Describe how transcriptomes are studied and discuss the applications of this type of research
7. Give a detailed description of the various levels of protein structure
8. Explain why amino acids underlie protein diversity
9. Outline how the meaning of each codon in the genetic code was elucidated
10. Describe the key features of the genetic code
11. Explain why the function of a protein is dependent on its amino acid sequence
12. List the major roles of proteins in living organisms and relate this diversity to the function of the genome

THE GENOME is a store of biological information but on its own it is unable to release that information to the cell. Utilization of the biological information requires the coordinated activity of enzymes and other proteins, which participate in a complex series of biochemical reactions referred to as genome expression. The details of genome expression are described in Part 3. Before reaching this detailed discussion, an overview of the key events involved in genome expression will be valuable, in order to establish a foundation of knowledge onto which the more comprehensive understanding can subsequently be built. This chapter provides that overview.

3.1. Genome Expression in Outline

The initial product of genome expression is the transcriptome, a collection of RNA molecules derived from those protein-coding genes whose biological information is required by the cell at a particular time (*Figure 3.1*). These RNA molecules direct synthesis of the final product of genome expression, the proteome, the cell's repertoire of proteins, which specifies the nature of the biochemical reactions that the cell is able to carry out.

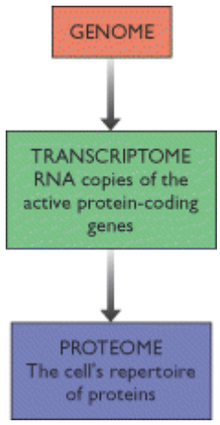

Figure 3.1. The genome, transcriptome and proteome

The transcriptome is constructed by the process called transcription, in which individual genes are copied into RNA molecules. Construction of the proteome involves translation of these RNA molecules into protein. Transcription and translation are important terms but it is unfortunate that the expression of individual genes is sometimes described simply as the two-step process 'DNA makes RNA makes protein' (*Figure 3.2A*). This is an inadequate over-simplification of the much more complex series of events involved in synthesis and maintenance of the transcriptome and proteome of even the simplest type of cell. In reality, genome expression comprises the following steps (*Figure 3.2B*):

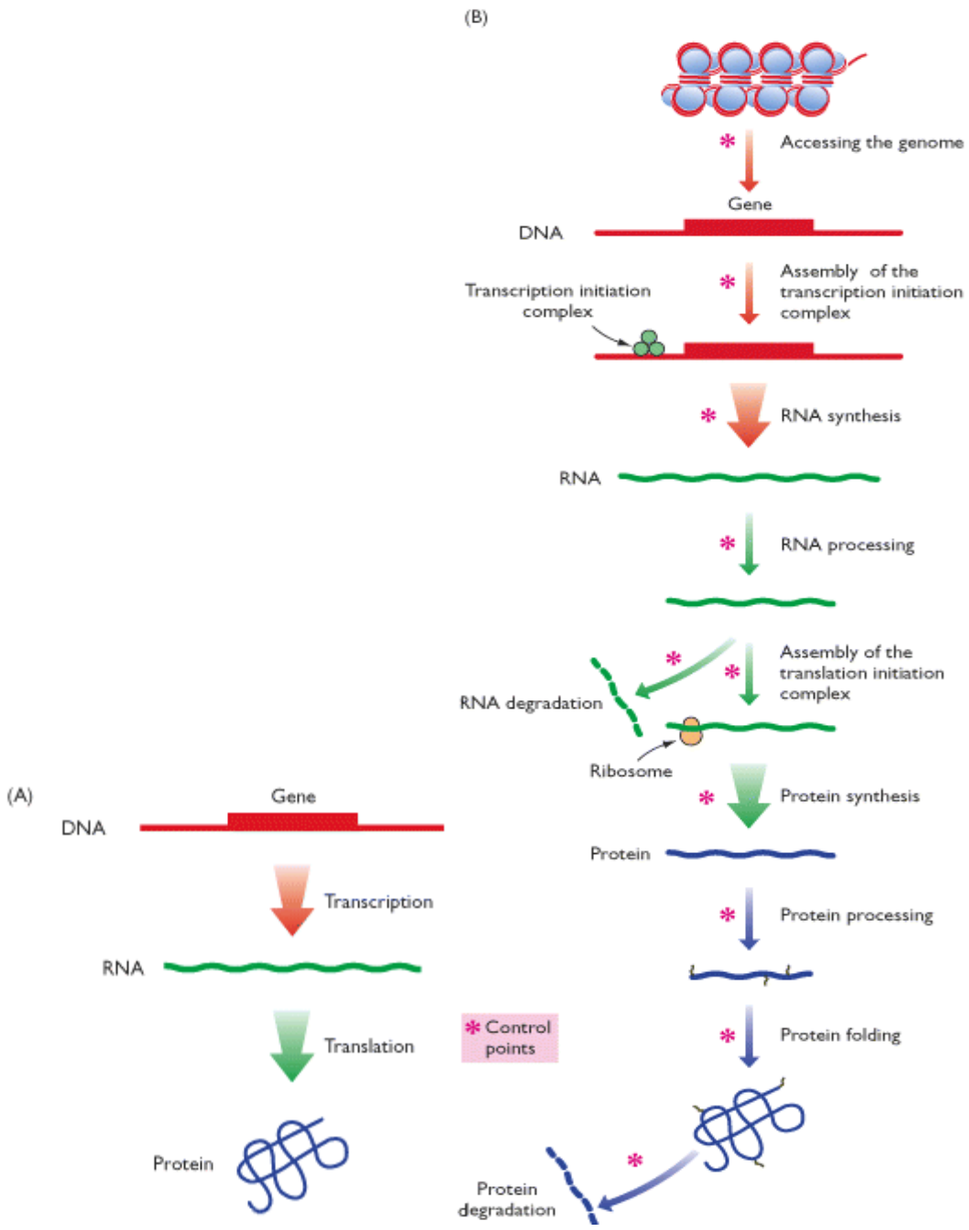

Figure 3.2. Two views of genome expression. (A) shows the old-fashioned depiction of gene expression, summarized as 'DNA makes RNA makes protein', the two steps being called transcription and translation. (B) gives a more accurate outline of the events involved in genome expression, especially in higher organisms. Note that these schemes apply only to protein-coding genes. Those genes that give rise to non-coding RNAs are transcribed and processed as shown but the RNAs are not translated

- **Accessing the genome.** This involves various processes that influence chromatin structure and nucleosome positioning in the parts of the genome that contain active genes, ensuring that these genes are accessible and are not buried deep within highly packaged parts of the chromosomes.
- **Assembly of the transcription initiation complex,** which comprises the set of proteins that work together to copy genes into RNA. Assembly of initiation complexes is a highly targeted process

because these complexes must be constructed at precise positions in the genome, adjacent to active genes, and nowhere else.

- **Synthesis of RNA,** during which the gene is transcribed into an RNA copy.
- **Processing of RNA** involves a series of alterations that are made to the sequence of the RNA molecule and to its chemical structure, and which must occur before the RNA molecules can be translated into protein or, in the case of non-coding RNA, before they can carry out their other functions in the cell.
- **RNA degradation** is the controlled turnover of RNA molecules. Degradation is not simply a means of getting rid of unwanted RNAs: it plays an active role in determining the make up of the transcriptome and hence is an integral step in genome expression.
- **Assembly of the translation initiation complex** occurs near the 5′ termini of coding RNA molecules, and is a prerequisite for translation of these molecules.
- **Protein synthesis** is the synthesis of a protein by translation of an RNA molecule.
- **Protein folding and protein processing** may occur together. Folding results in the protein taking up its correct three-dimensional configuration. Processing involves modification of the protein by addition of chemical groups and, for some proteins, removal of one or more segments of the protein.
- **Protein degradation** has an important influence on the composition of the proteome and, like RNA degradation, is an integral component of genome expression.

Genome expression is clearly much more complicated than ‘DNA makes RNA makes protein'. A particular weakness of this two-step interpretation is that it draws attention away from the points in the expression pathway at which the flow of information from genome to proteome can be regulated. Control mechanisms exist for regulating every one of the steps shown in *Figure 3.2B* , enabling the composition of the transcriptome and proteome to be altered in a rapid and precise manner, and allowing the cell to adjust its biochemical capabilities in response to changes in the extracellular environment and to signals received from other cells. As we will see in Chapter 12, these regulatory events underlie not only the functioning of individual cells but also the processes of differentiation and development.

3.2. The RNA Content of the Cell

A typical bacterium contains 0.05–0.10 pg of RNA, making up about 6% of its total weight. A mammalian cell, being much larger, contains more RNA, 20–30 pg in all, but this represents only 1% of the cell as a whole (Alberts *et al.*, 1994). It is important to appreciate that not all of this RNA constitutes the transcriptome. The latter is just the coding RNA - those molecules that have been transcribed from protein-coding genes and which are therefore capable of being translated into protein. Most of the cellular RNA does not fall into this category because it is non-coding. An understanding of the distinctive features of coding and non-coding RNA is therefore essential before we continue with our overview of genome expression.

3.2.1. Coding and non-coding RNA

The best way to understand the RNA content of a cell is to divide it into categories and subcategories depending on function. There are several ways of doing this, the most informative scheme being the one shown in *Figure 3.3* . The primary division is between coding RNA and non-coding RNA. The coding RNA comprises the transcriptome and is made up of just one class of molecule:

- Messenger RNAs (**mRNAs**), which are transcripts of protein-coding genes and hence are translated into protein in the latter stages of genome expression.

Messenger RNAs rarely make up more than 4% of the total RNA and are short-lived, being degraded soon after synthesis. Bacterial mRNAs have half-lives of no more than a few minutes and in eukaryotes most mRNAs are degraded a few hours after synthesis. This rapid turnover means that the composition of the transcriptome is not fixed and can quickly be restructured by changing the rate of synthesis of individual mRNAs.

The second type of RNA is non-coding. This is more diverse than the coding RNA and comprises transcripts with a number of different functions, all of which are performed by the RNA molecules themselves. In both prokaryotes and eukaryotes the two main types of non-coding RNA are:

- Ribosomal RNAs (**rRNAs**), which are the most abundant RNAs in the cell, making up over 80% of the total in actively dividing bacteria. These molecules are components of ribosomes, the structures on which protein synthesis takes place (Section 11.2).
- Transfer RNAs (**tRNAs**) are small molecules that are also involved in protein synthesis, carrying amino acids to the ribosome and ensuring that these are linked together in the order specified by the nucleotide sequence of the mRNA that is being translated (Section 11.1).

Ribosomal and tRNAs are present in the cells of all species. The other non-coding RNA types are more limited in their distribution (see *Figure 3.3*). Eukaryotes, for example, have a variety of short non-coding RNAs that are usually divided into three categories, the names indicating their primary locations in the cell:

- Small nuclear RNA (**snRNA**; also called U-RNA because these molecules are rich in uridine nucleotides), which is involved in mRNA processing (Section 10.1.3);
- Small nucleolar RNA (**snoRNA**), which plays a central role in the processing of rRNA molecules (Section 10.3.1);
- Small cytoplasmic RNA (**scRNA**), a diverse group including molecules with a range of functions, some understood and others still mysterious.

Bacteria and archaea also contain non-coding RNAs other than rRNA and tRNA but these molecules do not make up a substantial fraction of the total RNA. In bacteria they include one interesting RNA type, apparently present in most if not all species, called **transfer-messenger RNA (tmRNA)**, which looks like a tRNA attached to an mRNA, and which adds short peptide tags onto proteins that have been synthesized incorrectly, labeling them for immediate degradation (Muto *et al.*, 1998).

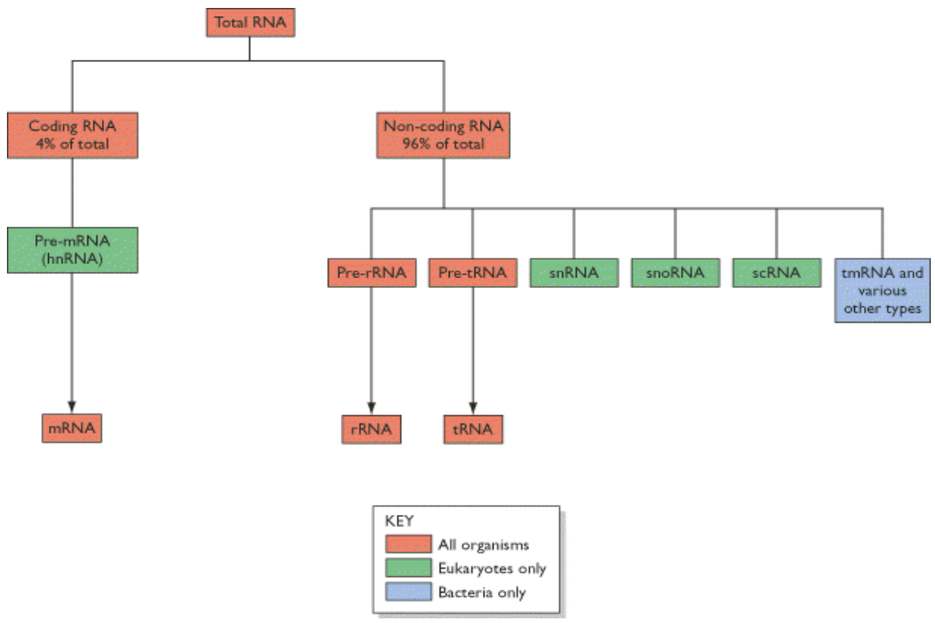

Figure 3.3. The RNA content of a cell. This scheme shows the types of RNA present in all organisms (eukaryotes, bacteria and archaea) and those categories found only in eukaryotic or bacterial cells. The non-coding RNAs of archaea have not yet been fully characterized and it is not clear which types are present in addition to rRNA and tRNA. For abbreviations, see the text.

3.2.2. Synthesis of RNA

The enzymes responsible for transcription of DNA into RNA are called DNA-dependent RNA polymerases. The name indicates that the enzymatic reaction that they catalyze results in polymerization of RNA from ribonucleotides, and occurs in a DNA-dependent manner, meaning that the sequence of nucleotides in a DNA template dictates the sequence of nucleotides in the RNA that is made (*Figure 3.4*). It is permissible to shorten the enzyme name to RNA polymerase, as the context in which the name is used means that there is rarely confusion with the RNA-dependent RNA polymerases that are involved in replication and expression of some virus genomes. Note that there are also template-independent RNA polymerases, including one - **poly(A) polymerase** - with an important role in genome expression (Section 10.1.2). The chemical basis of the template-dependent synthesis of RNA is illustrated in *Figure 3.5* . Ribonucleotides are added one after another to the growing 3′ end of the RNA transcript, the identity of each nucleotide being specified by the base-pairing rules: A base-pairs with T or U; G base-pairs with C. During each

nucleotide addition, the β- and γ-phosphates are removed from the incoming nucleotide, and the hydroxyl group is removed from the 3′-carbon of the nucleotide at the end of the chain, precisely the same as for DNA polymerization (see *Figure 1.8*).

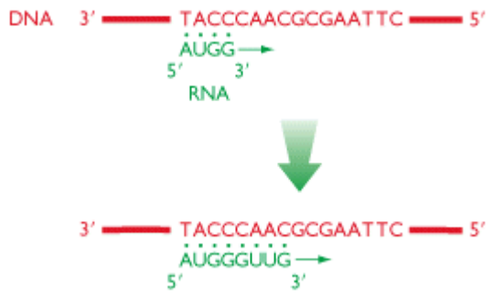

Figure 3.4. Template-dependent RNA synthesis. The RNA transcript is synthesized in the 5′→3′ direction, reading the DNA in the 3′→5′ direction, with the sequence of the transcript determined by base-pairing to the DNA template

RNA polymerase is the central component of the transcription initiation complex. Every time a gene is transcribed, a new complex has to be assembled immediately upstream of the gene. The initiation complexes are constructed at the appropriate positions, and not at random points within the genome, because their target sites are marked by specific nucleotide sequences called promoters, which are only found upstream of genes. In bacteria, promoters are directly recognized by the RNA polymerase enzyme, but in eukaryotes and archaea an intermediary DNA-binding protein is required, which attaches to the DNA and forms a platform to which the RNA polymerase binds (*Figure 3.6*). Promoters are described in detail in Section 9.2.2, which considers assembly of transcription initiation complexes.

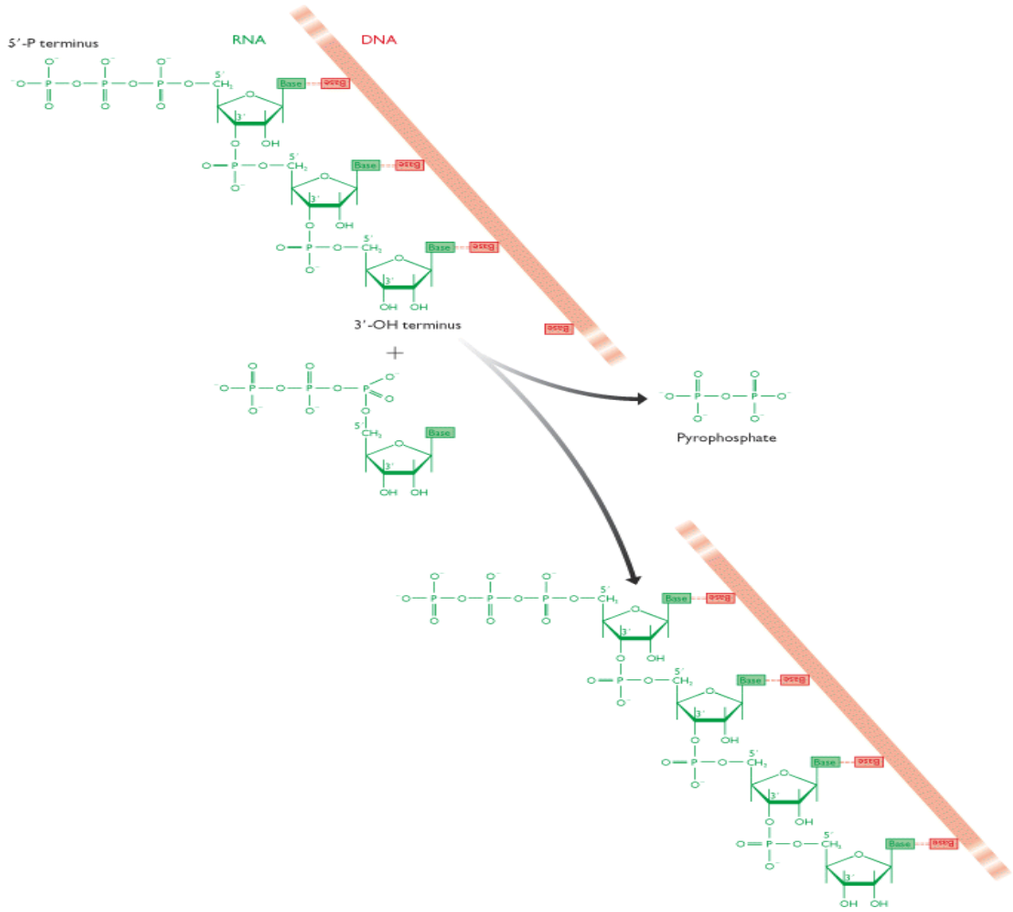

Figure 3.5. The chemical basis of RNA synthesis. Compare this reaction with polymerization of DNA, as illustrated in *Figure 1.8*

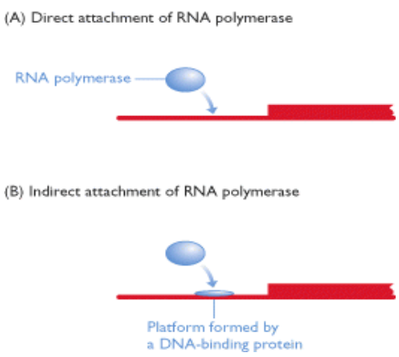

Figure 3.6. Two ways in which RNA polymerases bind to their promoters. (A) shows the direct recognition of the promoter by the RNA polymerase, as occurs in bacteria. (B) shows recognition of the promoter by a DNA-binding protein which forms a platform onto which the RNA polymerase binds. This indirect mechanism occurs with eukaryotic and archaeal RNA polymerases

Processing of precursor RNA

As well as the mature RNAs described above, cells also contain precursor molecules (see *Figure 3.3*). Many RNAs, especially in eukaryotes, are initially synthesized as precursor or pre-RNA, which has to be processed before it can carry out its function. The various processing events, all of which are described in detail in Chapter 10, include the following (*Figure 3.7*):

- End-modifications occur during the synthesis of eukaryotic and archaeal mRNAs, most of which have a single, unusual nucleotide called a cap attached at the 5′ end and a **poly(A) tail** attached to the 3′ end. The cap and poly(A) tails are involved in assembly of the translation initiation complex on these mRNAs (Section 10.2.2).
- Splicing is the removal of introns from a precursor RNA. Many eukaryotic protein-coding genes contain introns and these are copied when the gene is transcribed. The introns are removed from the pre-mRNA by cutting and joining reactions. Unspliced pre-mRNA forms the nuclear RNA fraction called heterogenous nuclear RNA (**hnRNA**). Some eukaryotic pre-rRNAs and pre-tRNAs also contain introns, as do some archaeal transcripts, but they are extremely rare in bacteria.
- **Cutting events** are particularly important in the processing of rRNA and tRNA, many of which are initially synthesized from transcription units that specify more than one molecule. The pre-rRNAs and pre-tRNAs must therefore be cut into pieces to produce the mature RNAs. This type of processing occurs in both prokaryotes and eukaryotes.
- **Chemical modifications** are made to rRNAs, tRNAs and mRNAs. The rRNAs and tRNAs of all organisms are modified by addition of new chemical groups, these groups being added to specific nucleotides within each RNA. Chemical modification of mRNA, called RNA editing, is seen in a diverse group of eukaryotes.

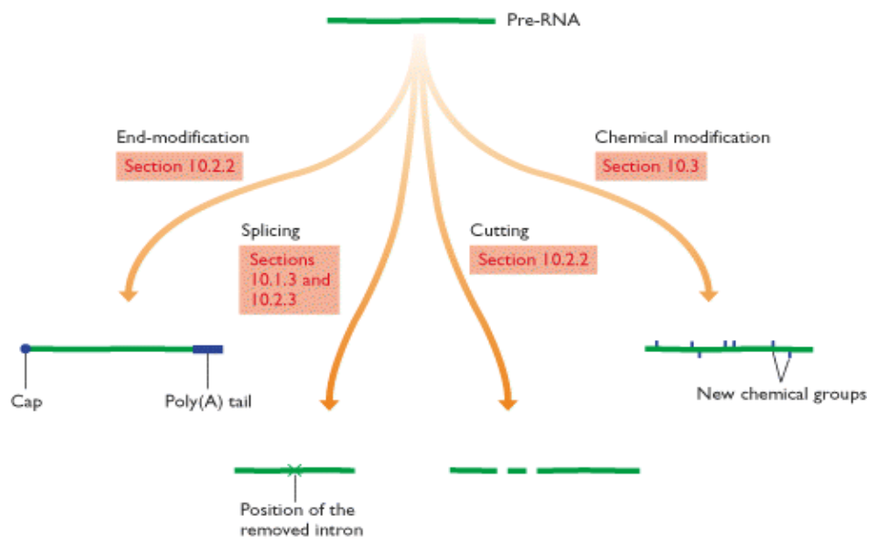

Figure 3.7. Schematic representation of the four types of RNA processing event. Not all events occur in all organisms - see the text for details.

Processing of mRNA has an important influence on the composition of the transcriptome. RNA editing, for example, can result in a single pre-mRNA being converted into two different mRNAs coding for quite distinct proteins (*Figure 3.8A*). This does not appear to be particularly common, but alternative splicing, in which one pre-mRNA gives rise to two or more mRNAs by assembly of different combinations of exons (*Figure 3.8B*), is fairly widespread. The mRNAs resulting from both editing and alternative splicing often display tissue specificity, the processing events increasing the coding capabilities of the genome without the requirement for an increased gene number.

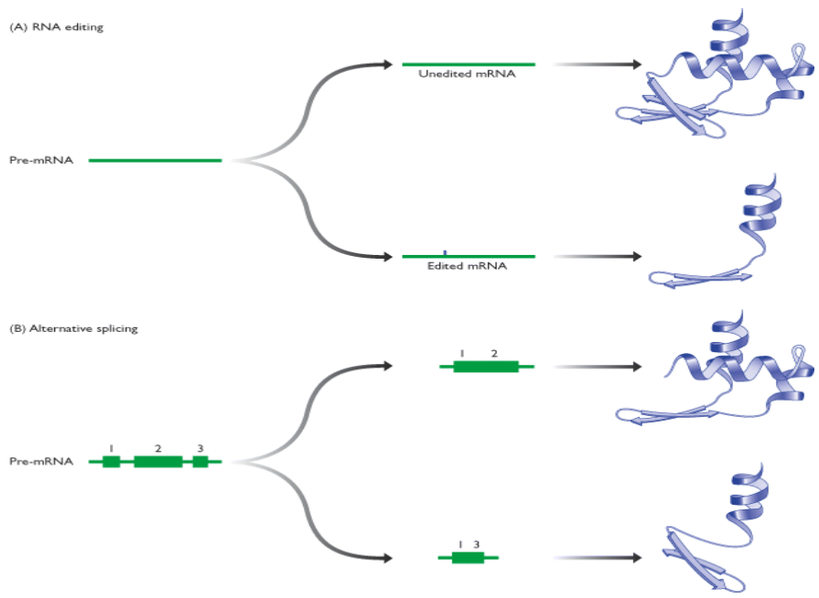

Figure 3.8. Some RNA processing events change the coding properties of an mRNA. (A) RNA editing can change the sequence of an mRNA, resulting in synthesis of a different protein. An example occurs with the human mRNA for apolipoprotein B, as shown in *Figure 10.29* . (B) Alternative splicing results in different combinations of exons becoming linked together, again resulting in different proteins being synthesized from the same pre-mRNA. *Figure 10.20* shows how alternative splicing underlies sex determination in *Drosophila*

3.2.3. The transcriptome

Although the transcriptome makes up less than 4% of the total cell RNA, it is the most significant component because it contains the coding RNAs that specify the composition of the proteome and hence determine the biochemical capacity of the cell. One important point to note is that the transcriptome is never synthesized *de novo*. Every cell receives part of its parent's transcriptome when it is first brought into existence by cell division, and maintains a transcriptome throughout its lifetime. Even quiescent cells in bacterial spores or in the seeds of plants have a transcriptome, although expression of that transcriptome into protein may be completely switched off. Transcription does not therefore result in *synthesis* of the transcriptome but instead *maintains* the transcriptome by replacing mRNAs that have been degraded, and brings about *changes* to the composition of the transcriptome via the switching on and off of different sets of genes.

Even in the simplest organisms such as bacteria and yeast, many genes are active at any one time. Transcriptomes are therefore complex, containing copies of hundreds, if not thousands, of different mRNAs. Usually, each mRNA makes up only a small fraction of the whole, with the most common type rarely contributing more than 1% of the total mRNA. Exceptions are those cells that have highly specialized biochemistries, which are reflected by transcriptomes in which one or a few mRNAs predominate. Developing wheat seeds are an example: these synthesize large amounts of the gliadin proteins, which accumulate in the dormant grain and provide a source of amino acids for the germinating seedling. Within the developing seeds, the gliadin mRNAs can make up as much as 30% of the transcriptomes of certain cells.

Perhaps surprisingly, it is relatively easy to determine the composition of a transcriptome, and to make comparisons between different transcriptomes, using the microarray and DNA chip technologies described in Section 7.3.1. To illustrate the types of analysis that are possible we will examine some of the recent research on the yeast and human transcriptomes.

Studies of the yeast transcriptome

With less than 6000 genes, the yeast *Saccharomyces cerevisiae* is ideally suited for transcriptome studies, and many of the pioneering projects have been carried out with this organism. One of the first discoveries was that, although mRNAs are being degraded and re-synthesized all the time, the composition of the yeast transcriptome undergoes very little change if the biochemical features of the environment remain constant (DeRisi *et al.*, 1997). When yeast is grown in a glucose-rich medium, which allows the cells to divide at their maximum rate, the transcriptome is almost completely stable, only 19 mRNAs displaying a greater than two-fold change in abundance over a period of 2 hours (*Figure 3.9*). Significant alterations to the transcriptome are seen only when the glucose in the growth medium becomes depleted, forcing the cells to switch from aerobic to anaerobic respiration. During this switch, the levels of over 700 mRNAs increase by a factor of two or more, and another 1000 mRNAs decline to less than half their original amount. The changing environment clearly results in a restructuring of the transcriptome to meet the new biochemical demands of the cell.

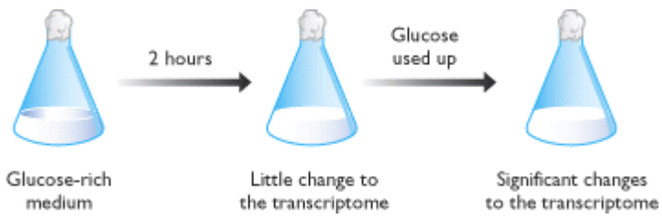

Figure 3.9. The yeast transcriptome is stable during growth of cells in a glucose-rich medium, but undergoes significant changes when the glucose is used up and the cells switch from aerobic to anaerobic respiration

The yeast transcriptome also undergoes restructuring during cellular differentiation. This has been established by studying sporulation (spore formation), which is induced by starvation and other stressful environmental conditions. The sporulation pathway can be divided into four stages - early, middle, mid-late and late - on the basis of the morphological and biochemical events that occur (*Figure 3.10*). Previous studies have shown, not unexpectedly, that each stage is characterized by expression of a different set of genes. Transcriptome studies have added to our understanding of the sporulation process in several ways (Chu *et al.*, 1998). Most significantly, the changes that occur to the composition of the transcriptome indicate that the early stage of sporulation can be subdivided into three distinct phases, called early (I), early (II) and early-middle. The levels of over 250 mRNAs increase significantly during early sporulation, and another 158 mRNAs increase specifically during the middle stage. A further 61 mRNAs increase in abundance during the mid-late period and five more during the late phase. There are also 600 mRNAs that decrease in abundance during sporulation, these presumably coding for proteins that are needed during vegetative growth but whose synthesis must be switched off when spores are being formed.

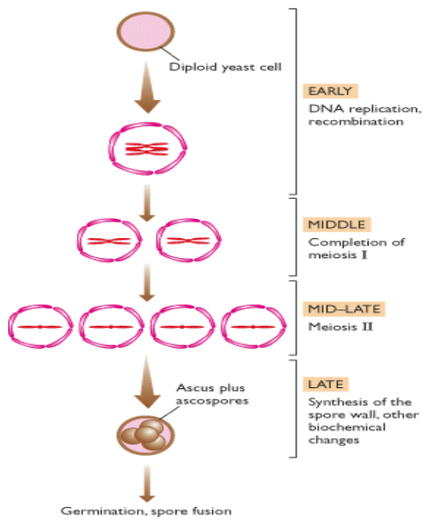

Figure 3.10. The sporulation pathway of *Saccharomyces cerevisiae*. The middle three drawings show the nuclear divisions that occur during sporulation. See *Figure 5.15* for details of the events involved in meiosis I and meiosis II.

This work on yeast sporulation is important for two reasons. First, by describing the changes in genome expression that occur during sporulation, the transcriptome analyses open the way to studies of the interactions between the genome and the environmental signals that trigger sporulation. Studies of this type, in a relatively simple organism such as yeast, act as an important model for the more complex developmental processes that operate in higher eukaryotes, including humans. Secondly, several of the mRNAs whose levels change significantly during sporulation are transcripts of genes whose functions were previously unknown. Transcriptome studies therefore help to annotate a genome sequence, aiding identification of genes whose roles in the genome have not been determined by other methods. We will return to this issue in Chapter 7.

The human transcriptome

With six times as many genes, the human transcriptome is substantially more complex than that of yeast, and studies of its composition are still in their infancy. Some interesting results have, nonetheless, been obtained. For example, the transcriptomes of eight different cell types have been mapped onto the draft human genome sequence (Caron *et al.*, 2001), resulting in global views of the pattern of gene expression along entire chromosomes (*Figure 3.11*). As well as providing a 'blueprint' for the transcriptomes of each of the eight cell types, this work has underlined the extent to which the transcriptome of a cancerous tissue differs from that of the normal version. Transcriptome restructuring as a result of cancer was first discovered several years ago, when it was shown that 289 mRNAs are present in significantly different amounts in the transcriptomes of normal colon epithelial cells compared with cancerous colon cells, and that about half of these mRNAs also display an altered abundance in pancreatic cancer cells (Zhang *et al.*, 1997). It is hoped that by understanding the differences between the transcriptomes of normal and cancerous cells it will be possible to devise new ways of treating the cancers.

Figure 3.11. Comparison of the transcriptomes of different types of human cell. The diagram shows human chromosome 11 aligned vertically. The bar charts indicate the expression levels in different cell types of the genes on this chromosome. The lengths of the blue bars are proportional to the extent of gene expression, and the red bars indicate genes whose expression levels are higher than can be illustrated on this scale. The box highlights significant differences between the transcriptomes of normal and cancerous breast tissue cells. Reprinted with permission from Caron *et al*. *Science*, **291**, 1289-1292. Copyright 2001 American Association for the Advancement of Science

Transcriptome studies also have applications in cancer diagnosis. The initial breakthrough in this respect came in 1999 when it was shown that the transcriptome of acute lymphoblastic leukemia cells is different from that of acute myeloid leukemia cells (Golub *et al.*, 1999). Twenty-seven lymphoblastic and eleven myeloid cancers were studied and, although all the transcriptomes were slightly different, the distinctions between the two types were sufficient for unambiguous identifications to be made. The significance of this work lies with the improved remission rates that are achievable if a cancer is identified accurately at an early stage, before clear morphological indicators are seen. This is not relevant with these two types of leukemia because these can be distinguished by non-genetic means, but it is important with other cancers such as non-Hodgkin lymphoma. The commonest version of this disease is called diffuse large B-cell lymphoma, and for many years it was thought that all tumors of this type were the same. Transcriptome studies changed this view and showed that B-cell lymphoma can be divided into two distinct subtypes (Alizadeh *et al.*, 2000). The distinctions between the transcriptomes of the two subtypes enable each one to be related to a different class of B cells, stimulating and directing the search for specific treatments that are tailored for each lymphoma.

3.3. The Protein Content of the Cell

The proteome is the final product of genome expression and comprises all the proteins present in a cell at a particular time. A 'typical' mammalian cell, for example a liver hepatocyte, is thought to contain 10 000–20 000 different proteins, about 8×10^9 individual molecules in all, representing approximately 0.5 ng of protein or 18–20% of the total cell weight (Alberts *et al.*, 1994; Lodish *et al.*, 2000). The copy numbers of individual proteins vary enormously, from less than 20 000 molecules per cell for the rarest types to 100 million copies for the commonest ones. Any protein that is present at a copy number of

greater than 50 000 per cell is considered to be relatively abundant, and in the average mammalian cell some 2000 proteins fall into this category. When the proteomes of different types of mammalian cell are examined, very few differences are seen among these abundant proteins, suggesting that most of them are **housekeeping** proteins which perform general biochemical activities that occur in all cells. The proteins that provide the cell with its specialized function are often quite rare, although there are exceptions such as the vast amounts of hemoglobin that are present only in red blood cells (Alberts *et al.*, 1994).

The proteome can be looked upon as the central link between the genome and the cell: it is, on the one hand, the culmination of genome expression and, on the other hand, the starting point for the biochemical activities that constitute cellular life (*Figure 3.12*). In order to comprehend how the proteome makes this connection we must first understand the structure of proteins.

Figure 3.12. The central role of the proteome

3.3.1. Protein structure

A protein**,** like a DNA molecule, is a linear unbranched polymer. In proteins the monomeric subunits are called amino acids (*Figure 3.13*) and the resulting polymers, or polypeptides, are rarely more than 2000 units in length. As with DNA, the key features of protein structure were determined in the first half of the 20th century, this phase of protein biochemistry culminating in the 1940s and early 1950s with the elucidation by Pauling and Corey of the major conformations or secondary structures taken up by polypeptides (Pauling *et al.*, 1951; Pauling and Corey, 1953). In recent years, interest has focused on how these secondary structures combine to produce the complex three-dimensional shapes of proteins.

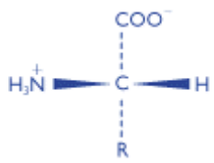

Figure 3.13. The general structure of an amino acid. All amino acids have the same general structure, comprising a central α-carbon attached to a hydrogen atom, a carboxyl group, an amino group and an R group. The R group is different for each amino acid (see *Figure 3.17*).

The four levels of protein structure

Proteins are traditionally looked upon as having four distinct levels of structure. These levels are hierarchical, the protein being built up stage by stage, with each level of structure depending on the one below it:

1. ***The primary structure*** of the protein is formed by joining amino acids into a polypeptide. The amino acids are linked by peptide bonds which are formed by a condensation reaction between the carboxyl group of one amino acid and the amino group of a second amino acid (*Figure 3.14*). In passing, note that, as with a polynucleotide, the two ends of the polypeptide are chemically distinct: one has a free amino group and is called the **amino**, **NH_2-**, or N terminus; the other has a free carboxyl group and is called the **carboxyl**, **COOH-**, or C terminus. The direction of the polypeptide can therefore be expressed as either N→C (left to right in *Figure 3.14*) or C→N (right to left in *Figure 3.14*).

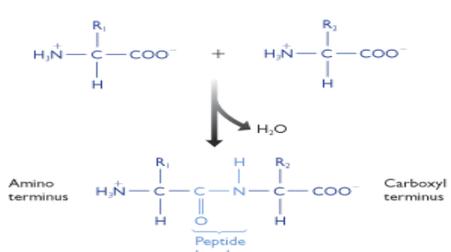

Figure 3.14. In polypeptides, amino acids are linked by peptide bonds. The drawing shows the chemical reaction that results in two amino acids becoming linked together by a peptide bond. The reaction is called a condensation because it results in elimination of water.

2. ***The secondary structure*** refers to the different conformations that can be taken up by the polypeptide. The two main types of secondary structure are the **α-helix** and **β-sheet** (*Figure 3.15*), both of which are stabilized by hydrogen bonds that form between different amino acids in the polypeptide. Most polypeptides are long enough to be folded into a series of secondary structures, one after another along the molecule.

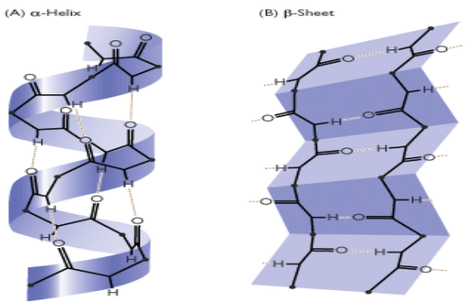

Figure 3.15. The two main secondary structural units found in proteins: (A) the α-helix, and (B) the β-sheet. The polypeptide chains are shown in outline with the positions of the α-carbons indicated by small dots. The R groups have been omitted for clarity. Each structure is stabilized by hydrogen bonds between the C=O and N-H groups of different peptide bonds. The β-sheet conformation that is shown is anti-parallel, the two chains running in opposite directions. Parallel β-sheets also occur

3. ***The tertiary structure*** results from folding the secondary structural components of the polypeptide into a three-dimensional configuration (*Figure 3.16*). The tertiary structure is stabilized by various chemical forces, notably hydrogen bonding between individual amino acids, and hydrophobic forces, which dictate that amino acids with non-polar (i.e. 'water-hating') side-groups must be shielded from water by embedding within the internal regions of the protein. There may also be covalent linkages called disulfide bridges between cysteine amino acids at various places in the polypeptide.

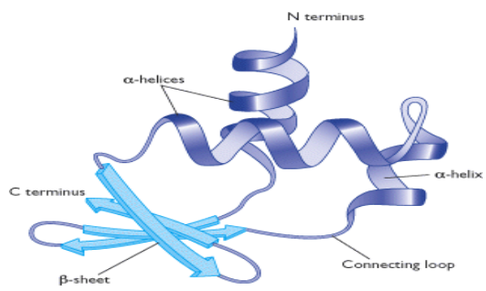

Figure 3.16. The tertiary structure of a protein. This imaginary protein structure comprises three α-helices, shown as coils, and a four-stranded β-sheet, indicated by the arrows. Redrawn from Turner *et al*. (1997)

4. ***The quaternary structure*** involves the association of two or more polypeptides, each folded into its tertiary structure, into a multi-subunit protein. Not all proteins form quaternary structures, but it is a feature of many proteins with complex functions, including several involved in genome expression. Some quaternary structures are held together by disulfide bridges between different polypeptides, but many proteins comprise looser associations of subunits stabilized by hydrogen bonding and hydrophobic effects, and can revert to their component polypeptides, or change their subunit composition, according to the functional requirements.

Amino acid diversity underlies protein diversity

Proteins are functionally diverse because the amino acids from which proteins are made are themselves chemically diverse. Different sequences of amino acids therefore result in different combinations of chemical reactivities, these combinations dictating not only the overall structure of the resulting protein but also the positioning on the surface of the structure of reactive groups that determine the chemical properties of the protein.

Amino acid diversity derives from the R group because this part is different in each amino acid and varies greatly in structure. Proteins are made up from a set of 20 amino acids (*Figure 3.17* ; *Table 3.1*). Some of these have R groups that are small, relatively simple structures such as a single hydrogen atom (in the amino acid called glycine) or a methyl group (alanine). Others are large complex aromatic side chains (phenylalanine, tryptophan and tyrosine). Most are uncharged, but two are negatively charged (aspartic acid and glutamic acid) and three are positively charged (arginine, histidine and lysine). Some are polar (e.g. glycine, serine and threonine), others are non-polar (e.g. alanine, leucine and valine).

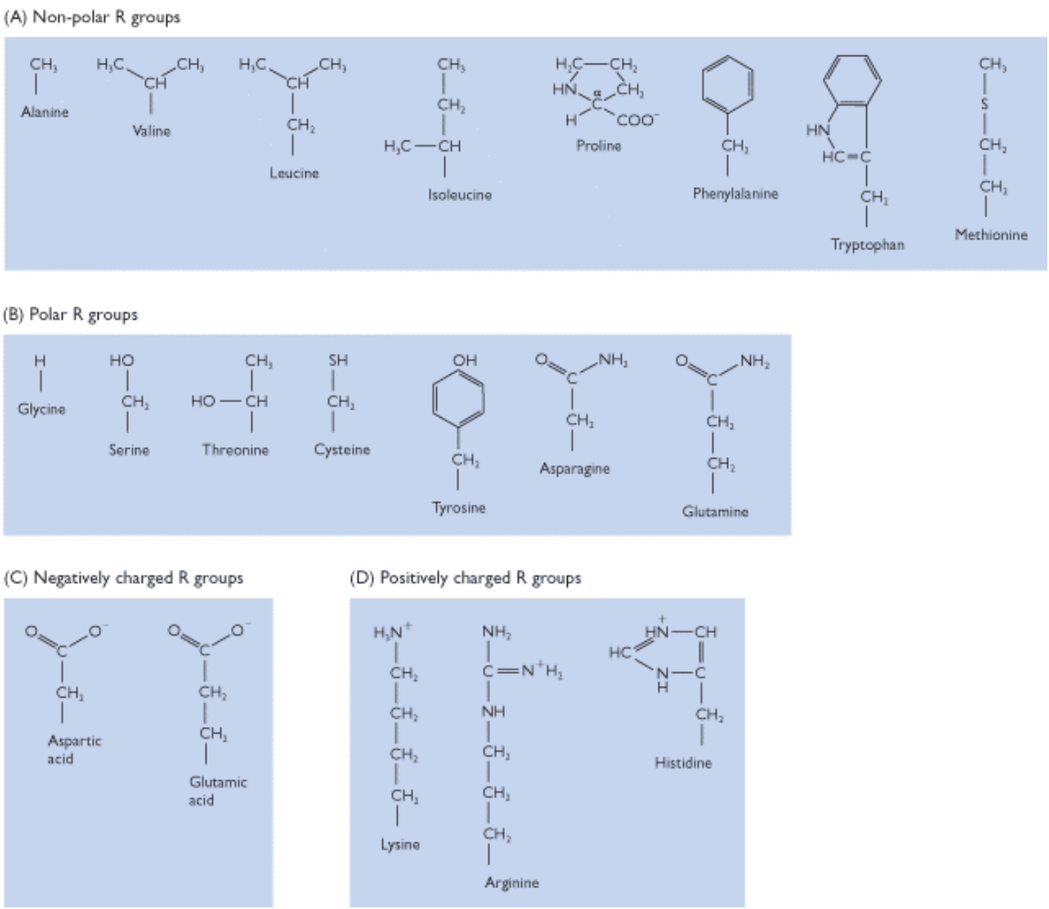

Figure 3.17. Amino acid R groups. These 20 amino acids are the ones that are conventionally looked upon as being specified by the genetic code (Section 3.3.2). The classification into non-polar, polar etc. is as described in Lehninger (1970)

The 20 amino acids shown in *Figure 3.17* are the ones that are conventionally looked on as being specified by the genetic code (Section 3.3.2). They are therefore the amino acids that are linked together when polypeptides are assembled during the protein-synthesis phase of genome expression. However, these 20 amino acids do not on their own represent the limit of the chemical diversity of proteins. The diversity is even greater because of two factors:

- At least one additional amino acid - selenocysteine (*Figure 3.18*) - can be inserted into a polypeptide chain during protein synthesis, its insertion directed by a modified reading of the genetic code (Section 3.3.2).
- During protein processing, some amino acids are modified by the addition of new chemical groups, for example by acetylation or phosphorylation, or by attachment of large side chains made up of sugar units (Section 11.3.3).

Proteins therefore have an immense amount of chemical variability, some of this directly specified by the genome, the remainder arising by protein processing.

Table 3.1. Amino acid abbreviations

	Abbreviation	
Amino acid	Three-letter	One-letter
Alanine	Ala	A
Arginine	Arg	R
Asparagine	Asn	N
Aspartic acid	Asp	D
Cysteine	Cys	C
Glutamic acid	Glu	E
Glutamine	Gln	Q
Glycine	Gly	G
Histidine	His	H
Isoleucine	Ile	I
Leucine	Leu	L
Lysine	Lys	K
Methionine	Met	M
Phenylalanine	Phe	F
Proline	Pro	P
Serine	Ser	S
Threonine	Thr	T
Tryptophan	Trp	W
Tyrosine	Tyr	Y
Valine	Val	V

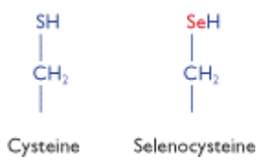

Figure 3.18. The R group of selenocysteine. Selenocysteine is the same as cysteine but with the sulfur replaced with a selenium atom

3.3.2. The link between the transcriptome and the proteome

The flow of information from DNA to RNA by transcription does not provide any conceptual difficulty. DNA and RNA polynucleotides have very similar structures and we can easily understand how an RNA copy of a gene can be made by template-dependent synthesis using the base-pairing rules with which we are familiar. The second phase of genome expression, during which the mRNA molecules of the transcriptome direct synthesis of proteins, is less easy to understand simply by considering the structures of the molecules that are involved. In the early 1950s, shortly after the double helix structure of DNA had been discovered, several molecular biologists attempted to devise ways in which amino acids could attach to mRNAs in an ordered fashion, but in all of these schemes at least some of the bonds had to be shorter or longer than was possible according to the laws of physical chemistry, and each idea was quietly dropped. Eventually, in 1957, Francis Crick cut a way through the confusion by predicting the existence of an adaptor molecule (Crick, 1990) that would form a bridge between the mRNA and the polypeptide being synthesized. Soon afterwards it was realized that the non-coding tRNAs were these adaptor molecules, and, once this fact had been established, a detailed understanding of the mechanism by which proteins are synthesized was quickly built up. We will examine this process in Section 11.1.

The other aspect of protein synthesis that interested molecular biologists in the 1950s and 1960s was the informational problem. This refers to the second important component of the link between the transcriptome and proteome: the genetic code which specifies how the nucleotide sequence of an mRNA is translated into the amino acid sequence of a protein.

The genetic code specifies how an mRNA sequence is translated into a polypeptide

It was recognized in the 1950s that a triplet genetic code - one in which each codeword or codon comprises three nucleotides - is required to account for all 20 amino acids found in proteins. A two-letter code would have only $4^2 = 16$ codons, which is not enough to account for all 20 amino acids, whereas a three-letter code would give $4^3 = 64$ codons. It was also assumed, as a working hypothesis, that mRNAs contain non-overlapping series of codons, and that these are colinear with the polypeptides that they encode (*Figure 3.19*). Although experimental proof of these assumptions was difficult to obtain, all three turned out to be correct. There are qualifications, such as the lack of strict colinearity displayed by many eukaryotic genes because of the presence of introns, but this complication was not appreciated until introns were discovered in 1977, long after the main work on the genetic code had been carried out. This work was completed in the mid-1960s when the meanings of all 64 codons were determined, partly by analysis of polypeptides resulting from translation of artificial mRNAs of known or predictable sequence in cell-free protein-synthesizing systems, and partly by determining which amino acids associated with which RNA sequences in an assay based on purified ribosomes, the protein-RNA complexes that carry out protein synthesis in the cell (Section 11.2). These experiments are described in more detail in Research Briefing 3.1.

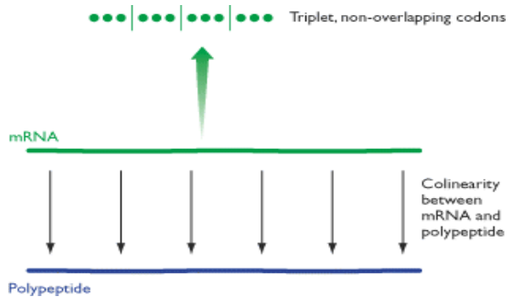

Figure 3.19. Early assumptions about the genetic code. It was assumed that the genetic code is triplet, that codons do not overlap, and that there is a colinear relationship between the sequences of an mRNA and the polypeptide it encodes

When the work on the genetic code was completed it was realized that the 64 codons fall into groups, the members of each group coding for the same amino acid (*Figure 3.20*). Only tryptophan and methionine have just a single codon each: all others are coded by two, three, four or six codons. This feature of the code is called degeneracy. The code also has four punctuation codons, which indicate the points within an mRNA where translation of the nucleotide sequence should start and finish (*Figure 3.21*). The initiation codon is usually 5′-AUG-3′, which also specifies methionine (so most newly synthesized polypeptides start with methionine), although with a few mRNAs other codons such as 5′-GUG-3′ and 5′-UUG-3′ are used. The three termination codons are 5′-UAG-3′, 5′-UAA-3′ and 5′-UGA-3′; these are sometimes called amber, opal and ochre, respectively, these being the whimsical names given to the original *Escherichia coli* mutants whose analysis led to their discovery.

UUU, UUC: phe UUA, UUG: leu	UCU, UCC, UCA, UCG: ser	UAU, UAC: tyr UAA, UAG: stop	UGU, UGC: cys UGA: stop UGG: trp
CUU, CUC, CUA, CUG: leu	CCU, CCC, CCA, CCG: pro	CAU, CAC: his CAA, CAG: gln	CGU, CGC, CGA, CGG: arg
AUU, AUC, AUA: ile AUG: met	ACU, ACC, ACA, ACG: thr	AAU, AAC: asn AAA, AAG: lys	AGU, AGC: ser AGA, AGG: arg
GUU, GUC, GUA, GUG: val	GCU, GCC, GCA, GCG: ala	GAU, GAC: asp GAA, GAG: glu	GGU, GGC, GGA, GGG: gly

Figure 3.20. The genetic code. See *Table 3.1* for the three-letter abbreviations of the amino acids.

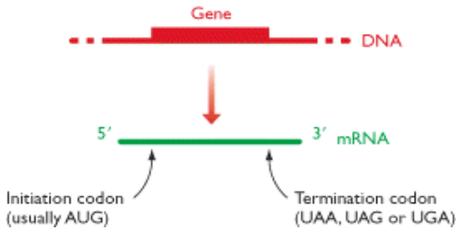

Figure 3.21. The positions of the punctuation codons in an mRNA

The genetic code is not universal

It was originally thought that the genetic code must be the same in all organisms. The argument was that, once established, it would be impossible for the code to change because giving a new meaning to any single codon would result in widespread disruption of the amino acid sequences of proteins. This reasoning seems sound, so it is surprising that, in reality, the genetic code is not universal. The code shown in *Figure 3.20* holds for the vast majority of genes in the vast majority of organisms, but deviations are widespread. In particular, mitochondrial genomes often use a non-standard code (*Table 3.2*). This was first discovered in 1979 by Frederick Sanger's group in Cambridge, UK, who found that several human mitochondrial mRNAs contain the sequence 5′-UGA-3′, which normally codes for termination, at internal positions where protein synthesis was not expected to stop. Comparisons with the amino acid sequences of the proteins coded by these mRNAs showed that 5′-UGA-3′ is a tryptophan codon in human mitochondria, and that this is just one of four code deviations in this particular genetic system. Mitochondrial genes in other organisms also display code deviations, although at least one of these - the use of 5′-CGG-3′ as a tryptophan codon in plant mitochondria - is probably corrected by RNA editing (Section 10.3.2) before translation occurs (Covello and Gray, 1989; Gualberto *et al.*, 1989).

Table 3.2. Examples of deviations from the standard genetic code

Organism	Codon	Should code for	Actually codes for
Mitochondrial genomes			
Mammals	UGA	Stop	Trp
	AGA, AGG	Arg	Stop
	AUA	Ile	Met
Drosophila	UGA	Stop	Trp
	AGA	Arg	Ser
	AUA	Ile	Met
Saccharomyces cerevisiae	UGA	Stop	Trp
	CUN	Leu	Thr
	AUA	Ile	Met
Fungi	UGA	Stop	Trp
Maize	CGG	Arg	Trp
Nuclear and prokaryotic genomes			
Several protozoa	UAA, UAG	Stop	Gln
Candida cylindracea	CUG	Leu	Ser
Micrococcus sp.	AGA	Arg	Stop
	AUA	Ile	Stop
Euplotes sp.	UGA	Stop	Cys
Mycoplasma sp.	UGA	Stop	Trp
	CGG	Arg	Stop
Context-dependent codon reassignments			
Various	UGA	Stop	Selenocysteine
Abbreviation: N, any nucleotide.			

Non-standard codes are also known in the nuclear genomes of lower eukaryotes. Often a modification is restricted to just a small group of organisms and frequently it involves reassignment of the termination codons (*Table 3.2*). Modifications are less common among prokaryotes but one example is known in *Mycoplasma* species. A more important type of code variation in nuclear genomes is context-dependent codon reassignment, which occurs when the protein to be synthesized contains selenocysteine. This applies to many organisms, both prokaryotes and eukaryotes, including humans, because selenoproteins

are widespread (see *Table 3.3*). Selenocysteine is coded by 5′-UGA-3′, which therefore has two meanings because it is still used as a termination codon in the organisms concerned (*Table 3.2*). The 5′-UGA-3′ codons that specify selenocysteine are distinguished from those that are true termination codons by the presence of a hairpin loop structure in the mRNA, positioned just downstream of the selenocysteine codon in prokaryotes and in the 3′ untranslated region (i.e. the part of the mRNA after the termination codon) in eukaryotes. Recognition of the codon requires interaction between the hairpin and a special protein that is involved in translation of these mRNAs (Low and Berry, 1996).

Table 3.3. Examples of proteins that contain selenocysteine

Protein	**Organism**
Prokaryotic enzymes	
Formate dehydrogenase	*Clostridium thermoaceticum, Clostridium thermoautotrophicum, Enterobacter aerogenes, Escherichia coli, Methanococcus vaniellii*
Glycine reductase	*Clostridium purinolyticum, Clostridium sticklandii*
NiFeSe hydrogenase	*Desulfomicrobium baculatum, Methanococcus voltae*
Eukaryotic enzymes	
Glutathione peroxidase	Human, cow, rat, mouse
Selenoprotein P	Human, cow, rat
Selenoprotein W	Rat
Type 1 deiodinase	Human, rat, mouse, dog
Type 2 deiodinase	Frog
Type 3 deiodinase	Human, rat, frog
See Low and Berry (1996).	

3.3.3. The link between the proteome and the biochemistry of the cell

How does the proteome convert the biological information that it has received from the genome into the biochemical capabilities of the cell? Two fundamental aspects of protein chemistry enable this final step in genome expression to be achieved. The first of these is the hierarchical nature of the four levels of protein structure, which provides a direct link between the amino acid sequence of a protein and its chemical properties. The second aspect is the multiplicity of the chemical properties that can be displayed by different proteins, this variability enabling proteins to carry out a huge range of different biochemical activities.

The amino acid sequence of a protein determines its function

The links between the different levels of protein structure are most clearly understood at the primary-to-secondary level. Because of the chemical properties of their R groups, certain amino acids are more commonly found in α-helices while others have a predisposition for β-sheets. A secondary structure therefore forms around a group of amino acids that favor that particular secondary structure and initiate its formation. It then extends to include adjacent amino acids that either favor the structure or have no strong disinclination towards it, and finally terminates when one or more blocking amino acids, which cannot participate in that particular type of structure, are reached (*Figure 3.22*). This process of nucleation, extension and delimitation is repeated along the polypeptide until each part of the chain has adopted its preferred secondary structure.

By identifying which amino acids are most frequently located in which secondary structures, and by studying the structures taken up by small polypeptides of known sequence, biochemists have been able to deduce rules for this level of protein folding, and, to a certain extent, can predict which secondary structures will be adopted by a polypeptide simply by examining its primary sequence (Barton, 1995). It is less easy to predict the outcomes of the next two stages of protein folding, which result in the secondary structural units becoming arranged into the tertiary structure, and tertiary units associating to form quaternary multi-subunit structures. The tertiary structures of most proteins are made up of two or more structural domains, possibly with little interaction between them; these domains are thought to fold independently of one another. Understanding how this occurs is complicated by the fact that many domains include secondary structural units from quite different regions of a polypeptide. But the difficulty in identifying rules for predicting tertiary and quaternary structures does not detract from the fact that these higher levels of structure are determined by the amino acid sequence of the polypeptide.

This is illustrated by the ability of proteins that have been unfolded in the test tube, for example by treatment with urea, to refold spontaneously into their correct structures when the treatment is reversed (Section 11.3.1).

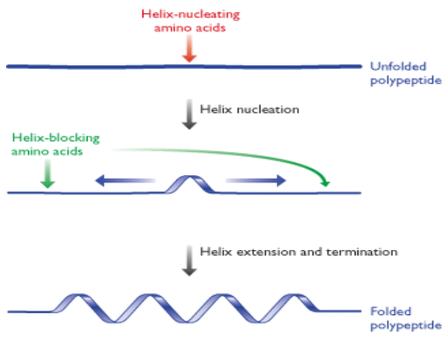

Figure 3.22. Formation of a secondary structure in a polypeptide. An α-helix is shown nucleating at a position containing amino acids that favor helix formation, and extending in either direction until groups of amino acids that block helix formation are reached.

Spontaneous refolding does not occur with all proteins and is particularly difficult to achieve with larger ones, the problem appearing to be that the protein can adopt alternative partially folded structures at various stages of the folding process, only one of which leads to the correctly folded tertiary configuration. If the protein makes the wrong 'choice' it can end up at a dead end in which it is partially folded in an incorrect manner but from which it cannot escape. In cells, proteins called molecular chaperones aid the folding of other proteins, probably by reducing the likelihood that the protein being folded adopts the wrong intermediate structure (Section 11.3.1). The existence of alternative folding pathways, and of proteins that aid the folding of other proteins, complicates the premise that a protein's folded structure is dictated by its amino acid sequence, but the premise still holds. The role of molecular chaperones is not to impose a new structure on a protein, but merely to increase the efficiency of the protein's natural, sequence-directed, folding pathway.

The multiplicity of protein function

The biological information encoded by the genome finds its final expression in a protein whose biological properties are determined by the spatial arrangement of chemical groups on its surface and within its folded structure. By specifying proteins of different types, the genome is able to construct and maintain a proteome whose overall biological properties form the underlying basis of life. The proteome can play this role because of the huge diversity of protein structures that can be formed, the diversity enabling proteins to carry out a variety of biological functions. These functions include the following:

- ***Biochemical catalysis*** is the role of the special type of proteins called enzymes. The central metabolic pathways, which provide the cell with energy, are catalyzed by enzymes, as are the biosynthetic processes that result in construction of nucleic acids, proteins, carbohydrates and lipids. Biochemical catalysis also drives genome expression through the activities of enzymes such as RNA polymerase.
- ***Structure***, which at the cellular level is determined by the proteins that make up the cytoskeleton, is also the primary function of some extracellular proteins. An example is collagen, which is an important component of bones and tendons.
- ***Movement*** is conferred by contractile proteins, of which actin and myosin in cytoskeletal fibers are the best known examples.
- ***Transport*** of materials around the body is an important protein activity: for example, hemoglobin transports oxygen in the bloodstream, and serum albumin transports fatty acids.
- ***Regulation*** of cellular processes is mediated by signaling proteins such as STATs (Section 12.1.2) and by proteins such as activators that bind to the genome and influence the expression levels of individual genes and groups of genes (Section 9.3.2). The activities of groups of cells are regulated and coordinated by extracellular hormones and cytokines, many of which are proteins (e.g. insulin, the hormone that controls blood sugar levels, and the interleukins, a group of cytokines that regulate cell division and differentiation).
- ***Protection*** of the body and of individual cells is the function of a range of proteins, including the antibodies, and those proteins involved in the blood clotting response.
- ***Storage*** functions are performed by proteins such as ferritin, which acts as an iron store in the liver, and the gliadins, which store amino acids in dormant wheat seeds.

This multiplicity of protein function provides the proteome with its ability to convert the blueprint contained in the genome into the essential features of the life process.